This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2428999, IEEE Transactions on Biomedical Engineering

A Stereo Music Pre-Processing Scheme for Cochlear Implant Users

Wim Buyens, Bas van Dijk, Jan Wouters, and Marc Moonen

***Abstract*— *Objective*: Listening to music is still one of the more challenging aspects of using a cochlear implant (CI) for most users. Simple musical structures, a clear rhythm/beat and lyrics that are easy to follow are among the top factors contributing to music appreciation for CI users. Modifying the audio mix of complex music potentially improves music enjoyment in CI users. *Methods:* A stereo music pre-processing scheme is described in which vocals, drums and bass are emphasized based on the representation of the harmonic and the percussive components in the input spectrogram, combined with the spatial allocation of instruments in typical stereo recordings. The scheme is assessed with post-lingually deafened CI subjects (N=7) using pop/rock music excerpts with different complexity levels. *Results:* The scheme is capable of modifying relative instrument level settings, with the aim of improving music appreciation in CI users, and allows individual preference adjustments. The assessment with CI subjects confirms the preference for more emphasis on vocals, drums and bass as offered by the pre-processing scheme, especially for songs with higher complexity. *Conclusion:* The stereo music pre-processing scheme has the potential to improve music enjoyment in CI users by modifying the audio mix in widespread (stereo) music recordings. *Significance:* Since music enjoyment in CI users is generally poor, this scheme can assist the music listening experience of CI users as a training or rehabilitation tool.**

***Index Terms*— cochlear implants, music processing, sound separation**

I. Introduction

A cochlear implant (CI) is a medical device enabling people with severe-to-profound sensorineural hearing loss to perceive sounds by electrically stimulating the auditory nerve using an electrode array implanted in the cochlea [1]. This type of hearing loss is mostly caused by malfunctions in the hair cells of the cochlea and can be congenital or acquired after birth. Although CI users reach good speech understanding in quiet surroundings, music perception and appreciation generally remain poor [2]. Simple musical structures, a clear rhythm/beat and lyrics that are easy to follow were reported amongst the top factors contributing to music appreciation in CI users [3]. A negative correlation was found between (subjective) complexity and appreciation, studied with pop, country and classical music [4]. CI users were asked to rate complexity and appraisal for different music excerpts on a scale from 0 to 100. Classical music was rated as more complex than pop and country music. Several plausible explanations were provided including the presence of simple musical structures and lyrics in pop and country music. Since CIs were mainly developed for transmitting speech sounds, the presence of lyrics may make it easier for CI users to follow the sequence of events in complex music. In addition, both pop and country music often contain a strong, simple beat which is well encoded in the electrical stimulation pattern of current CIs. Moreover, the performance for CI subjects on rhythmic pattern perception tasks is nearly as good as for normal hearing (NH) subjects (e.g. [2] [5] [6]). On the other hand, the accurate transmission of spectral and fine-structure information (as is often the case in instrumental classical music) remains challenging due to channel interactions, limited number of stimulation channels and limited dynamic range.

The preference for clear vocals and a strong rhythm/beat in CI users was demonstrated before by modifying relative instrument level settings in pop music [7]. A significant difference in preference rating scores was found between NH and CI subjects. For the pop songs provided, CI subjects preferred an audio mix with higher vocals-to-instruments ratio compared to NH subjects. In addition, given an audio mix with clear vocals and attenuated instruments, CI subjects preferred the drums/bass track to be louder than the other instrument tracks. Although individual differences occurred across subjects, the potential for improving music appraisal by modifying the audio mix is apparent. The relative instrument level settings were modified by altering the levels of the different, separately recorded, instrument tracks, which are, however, not widely available for most music. To accomplish the same modification in relative instrument levels with mono or stereo music recordings, a specific signal processing scheme is needed. By using sound source separation techniques, vocals, drums and bass can be separated out, and

This work is supported by a Baekeland PhD grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT090274) and Cochlear Technology Centre Belgium.

W. Buyens is with Cochlear Technology Centre Belgium, Schaliënhoevedreef 20 I, 2800 Mechelen, Belgium, with KU Leuven, Department of Neurosciences (ExpORL), O&N2, Herestraat 49 bus 721, 3000 Leuven and with KU Leuven, Department of Electrical Engineering (ESAT-STADIUS), Kasteelpark Arenberg 10, 3001 Heverlee. (e-mail: wim.buyens@esat.kuleuven.be)

B. van Dijk is with Cochlear Technology Centre Belgium.

J. Wouters is with KU Leuven, Department of Neurosciences (ExpORL).

M. Moonen is with KU Leuven, Department of Electrical Engineering (ESAT-STADIUS).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2428999, IEEE Transactions on Biomedical Engineering

then the residual signal is mixed in at a different level to provide the output of this specific signal processing scheme. Several approaches have been studied to tackle the sound source separation problem, and can be divided in two main categories: single-channel methods and multi-channel methods [8]. The approaches used in single-channel sound source separation can be categorized roughly into model-based inference, unsupervised learning and psycho-acoustically motivated methods. Mostly, a combination of these approaches is used in practice. In model-based inference a parametric model of the sound sources to be separated is employed in which the model parameters are estimated from the observed mixture signal. The sinusoids plus noise model is the most commonly used model, introduced in music signal processing by Smith and Serra [9] [10] [11]. Unsupervised learning methods use a simple non-parametric model and estimate the sound source characteristics from the data based on Independent Component Analysis (ICA), Non-Negative Matrix Factorization (NMF) or Sparse Coding. Uhle et al. applied ICA to the spectrogram, and classified the extracted independent components into a harmonic and a percussive group based on features like percussiveness, noise-likeness, etc. [12]. Helen and Virtanen utilized NMF for decomposing the spectrogram into elementary patterns and classified them by a pre-trained Support Vector Machine (SVM) [13]. Sparse coding methods represent a mixture signal in terms of a small number of active elements chosen from a larger set [14], and were used in the analysis of music signals by Abdallah and Plumbley [15] and Virtanen [16]. Although good performance is demonstrated for many different signals, the aforementioned approaches often have their limitations. First, some sound sources are hard to model with a few fixed spectral templates and in practice require careful tuning of sophisticated models [17] [18]. Second, the typical assumption that different sources are characterized by different sets of spectra may not be realistic in many cases in music signals.

Instead of decomposing the sound sources as a combination of fixed patterns, in psycho-acoustically motivated methods the elementary time-frequency components of the incoming signal are categorized into their respective sound sources based on association cues such as spectral proximity, harmonic concordance, synchronous changes and spatial proximity [19]. A simple and fast algorithm to perform harmonic/percussive sound separation is based on the "anisotropic smoothness" of the harmonic and percussive components in the spectrogram [20]. Harmonic components appear smooth in the temporal direction, whereas percussive components appear smooth in the frequency direction in the spectrogram. A (mono) music pre-processing scheme for CI users based on harmonic/percussive sound separation was described in [21] for a single-channel input. This scheme emphasized vocals and drums in complex pop music in order to improve music appreciation in CI users.

In contrast to single-channel methods, multi-channel methods can take advantage of the availability of spatial information, e.g., from recordings with multiple microphones placed at different positions, enabling acoustic beamforming or blind separation of convolutive mixtures to recover the sound sources. The typical karaoke problem to remove vocals from background music also exploits the spatial information of stereo recordings [22].

In this paper, a stereo music pre-processing scheme for CI users is described, which is a stereo extension of the mono music pre-processing scheme from [21] with improved performance based on exploiting the spatial information in stereo recordings. The building blocks for this stereo music pre-processing scheme are described in detail in section II. The evaluation of the scheme with CI users and pop music excerpts is presented in section III and IV.

II. Stereo Music Pre-Processing Scheme

The stereo music pre-processing scheme, which performs vocals, drums and bass enhancement on complex stereo music, is shown in Figure 1. The scheme contains a "Vocals & Drums Extraction" block applied on the input spectrogram and a "Stereo Binary Mask" block to promote components that are located in the center of the stereo image. The output of the scheme contains the extracted signal (indicated as "Vocals, Drums, Bass") mixed together with an attenuated version of the residual signal (indicated as "Other").

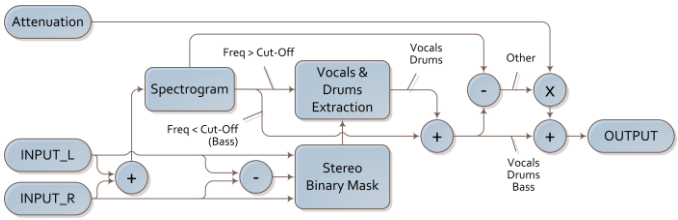

Fig. 1. Schematic of the stereo music pre-processing scheme for CI users which is enhancing vocals/drums/bass while attenuating the 'Other' instruments with parameter 'Attenuation'. It is based on "Vocals & Drums Extraction" of the input spectrogram (Input_L + Input_R with frequency > cut-off frequency) and a "Stereo Binary Mask" to exploit the spatial information in stereo recordings (based on Input_L, Input_R, Input_L-Input_R).

In paragraph A, the "Vocals & Drums Extraction" is described in detail, performing vocals and drums enhancement on a mono input signal which is – in the case of a stereo input – the sum of the left and the right channel. The processing of the bass frequencies is explained in paragraph B. In paragraph C the typical mixing settings in stereo recordings are explained and the inclusion of the stereo mixing property is described as a constraint in the optimization problem with the "Stereo Binary Mask". Paragraph D concludes with the computation of the output of the stereo music pre-processing scheme.

A. *Vocals & Drums Extraction*

The "Vocals & Drums Extraction" block is adopted from the mono music pre-processing scheme for CI users in [21]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2428999, IEEE Transactions on Biomedical Engineering

and is based on the harmonic/percussive sound separation (HPSS) in [20], which separates harmonic (H) and percussive (P) components by exploiting the "anisotropic smoothness" of these components in the spectrogram. "Anisotropic smoothness" is defined based on partial differentials of the spectrogram in the temporal or the frequency direction: harmonic components are "smooth in the temporal direction" because they are sustained and periodic; percussive components are "smooth in the frequency direction" because they are instantaneous and aperiodic [23]. The input spectrogram $W_{\tau,\omega} = STFT(w(t))$, which is calculated using the short-time Fourier transform (STFT) and a Hamming window, is decomposed into the harmonic components $H_{\tau,\omega}$ and the percussive components $P_{\tau,\omega}$. The indices τ and ω represent time and frequency, respectively. The L_2 norms of the spectrogram gradients are used as a metric for the anisotropic smoothness, that is, $H_{\tau,\omega}$ and $P_{\tau,\omega}$ are found by minimizing:

$$J(\boldsymbol{H},\boldsymbol{P}) = \frac{1}{2\sigma_H^2}\sum_{\tau,\omega}\left(H_{\tau-1,\omega} - H_{\tau,\omega}\right)^2 + \frac{1}{2\sigma_P^2}\sum_{\tau,\omega}\left(P_{\tau,\omega-1} - P_{\tau,\omega}\right)^2 \quad (1)$$

under the constraint of

$$H_{\tau,\omega}{}^2 + P_{\tau,\omega}{}^2 = \left|W_{\tau,\omega}\right|^2 \quad (2)$$

$$H_{\tau,\omega} \geq 0,\ P_{\tau,\omega} \geq 0 \quad (3)$$

where $\boldsymbol{H}$ and $\boldsymbol{P}$ are sets of $H_{\tau,\omega}$ and $P_{\tau,\omega}$, respectively, and σ_H and σ_P are parameters to control the weights of the horizontal and vertical smoothness. This optimization problem can be solved numerically, which results in the following iteration formulae [24]:

$$H_{\tau,\omega}^{2\ (j+1)} = \frac{\alpha_{\tau,\omega}{}^{(j)}\left|W_{\tau,\omega}\right|^2}{\left(\alpha_{\tau,\omega}{}^{(j)} + \beta_{\tau,\omega}{}^{(j)}\right)} \quad (4)$$

$$P_{\tau,\omega}^{2\ (j+1)} = \frac{\beta_{\tau,\omega}{}^{(j)}\left|W_{\tau,\omega}\right|^2}{\left(\alpha_{\tau,\omega}{}^{(j)} + \beta_{\tau,\omega}{}^{(j)}\right)} \quad (5)$$

where j is the iteration index, and

$$\alpha_{\tau,\omega}{}^{(j)} = \left(H_{\tau+1,\omega}{}^{(j)} + H_{\tau-1,\omega}{}^{(j)}\right)^2 \quad (6)$$

$$\beta_{\tau,\omega}{}^{(j)} = \kappa^2\left(P_{\tau,\omega+1}{}^{(j)} + P_{\tau,\omega-1}{}^{(j)}\right)^2 \quad (7)$$

$$\kappa = \frac{\sigma_H^2}{\sigma_P^2} \quad (8)$$

The tunable parameter κ is optimized to maximize the separation of vocals and drums from the other instruments by using (16) and the multi-track recordings (vocals, drums, bass, guitar, piano) used in [7]. From the estimated spectrograms $H_{\tau,\omega}$ and $P_{\tau,\omega}$, a time-frequency mask is defined which is then used to estimate the corresponding waveforms h(t) and p(t). From the considered time-frequency masks, the binary mask has been found more effective to improve the separation performance compared to the Wiener mask or not applying any mask [24]. The binary mask ($BM_{\tau,\omega}$) is defined as:

$$BM_{\tau,\omega} = 1 \qquad P_{\tau,\omega} > H_{\tau,\omega} \quad (9)$$

$$= 0 \qquad P_{\tau,\omega} \leq H_{\tau,\omega}$$

Using this binary mask, the P- and H-components are computed from the input spectrogram as:

$$p(t) = ISTFT\left(BM_{\tau,\omega}\,.\,W_{\tau,\omega}\right) \quad (10)$$

$$h(t) = ISTFT\left(\left(1 - BM_{\tau,\omega}\right).W_{\tau,\omega}\right) \quad (11)$$

Figure 2 shows the separation performance versus the number of iterations with and without binary mask for the multi-track recordings used in [7]. The benefit of the binary mask on the separation performance is clearly seen. The separation performance is defined as the signal-to-noise ratio of the P-components $p(t)$, in which the signal (S) consists of vocals and drums and the noise (N) represents the other instruments. Since multi-track recordings are available from the songs used in [7], the signal-to-noise ratio can be calculated at both the input and the output as:

$$SNR = 20 * log_{10}\left(\frac{rms(S)}{rms(N)}\right) \quad (12)$$

To indicate the presence of the different instruments in the P-components, the energy ratio for the different tracks is used, which is calculated as:

$$r_p^i = \frac{E_p^i}{E_h^i + E_p^i} \quad (13)$$

where

$$E_p^i = < f_i(t), p(t) > \quad (14)$$

$$E_h^i = < f_i(t), h(t) > \quad (15)$$

in which <> represents the cross-correlation operation with time-lag zero and $f_i(t)$ the signal of track i. Therefore, the separation of vocals and drums from the other instruments can also be defined as the difference (δ) in energy ratio of the P-components for vocals/drums and bass/guitar/piano:

$$\delta = \left(r_p^{vocals/drums}\right) - \left(r_p^{bass/guitar/piano}\right) \quad (16)$$

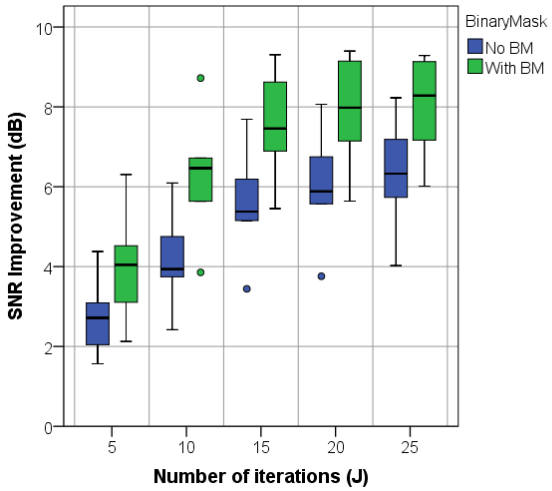

Fig. 2. Boxplot with SNR improvement (dB) of the P-components with vocals/drums versus the other instruments for the multi-track recordings used in [7] as a function of the number of iterations (J) with and without applying the binary mask (BM) from (9). The window length of the STFT used in this graph is 185 ms.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2428999, IEEE Transactions on Biomedical Engineering

The separation results from the "Vocals & Drums Extraction" block show that in addition to the drums, also the vocals can indeed be extracted as P-components based on their representation in the input spectrogram. "Temporally-variable" sounds are contrasted to "temporally-stable" sounds. The "temporally-variable" sounds (such as vocal tones) contain 4-8 Hz quasi-periodic vibrations of the fundamental frequencies (F0s) and do not sustain for a long time. On the other hand, the "temporally-stable" sounds (such as chord tones) contain very few fluctuations and are maintained stationary for a while [23]. Adjusting the time-frequency resolution of the STFT results in a different classification for the temporally-variable components. For a STFT with long time window (100-500 ms) percussive and temporally-variable sounds appear "smooth" in the frequency direction (P-components), whereas for a STFT with short time window (30 ms) temporally-stable sounds as well as temporally-variable sounds appear "smooth" in the temporal direction (H-components). This is illustrated in Figure 3 for a typical pop song with vocals, drums, bass, guitar and piano. The energy ratio of the P-components calculated with (13) is shown for every instrument as a function of the window length. Instruments with high energy ratio (close to one) appear as P-components, whereas instruments with low energy ratio (close to zero) appear as H-components. Figure 3 shows that the vocals appear in the H-components if a small STFT window length is adopted, whereas with a large STFT window length, the vocals appear in the P-components. A window length above 185 ms is good for vocals/drums separation, but the longer the window the more latency. Hence we choose in the "Vocals & Drums Extraction" block (Figure 1) a window length of 185 ms, resulting in the classification of temporally-variable components (such as vocals) as P-components. The small distortions on the vocals introduced by the algorithm are hardly noticeable for CI users. In speech, similar distortions introduced by a binary mask have been studied in [25], showing no degradation in quality for CI users, as opposed to NH listeners, but a significant improvement in speech understanding in noise.

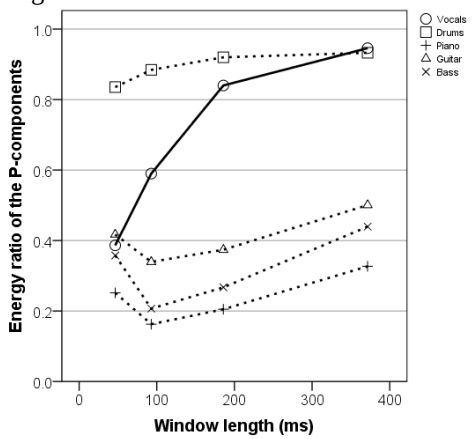

Fig. 3. Energy ratio of the P-components (j=15) for the different tracks of a typical pop song as a function of the STFT window length. The vocals (straight line) are separated as P-components with high STFT window length and as H-components with low STFT window length.

The impact of the proposed settings of the "Vocals & Drums Extraction" on other instruments is shown in Figure 4. The energy ratio of the P-components is computed for different instruments. An energy ratio close to zero means the instrument has been removed from the P-components, whereas an energy ratio close to one means the instrument remains in the P-components. Instrument samples are collected from the instrument recognition tests MACarena [26] and UW-CAMP [27][28]. The MACarena samples consist of the first 8 bars of a traditional Swedish folksong played by professional musicians on different instruments. The UW-CAMP samples consist of a 5-note melodic sequence played on different instruments with the same detached articulation. The majority of the instruments show low energy ratio and are thus removed from the P-components with the "Vocals & Drums Extraction". However, the flute and the violin from the MACarena test show a high energy ratio value due to the very strong *vibrato* in the music samples. This contrasts with the UW-CAMP flute and violin which have a low energy ratio value and no *vibrato*. The pitched percussion instruments (guitar and piano) are not completely removed in the P-components because of the clear percussive onset, except for the MACarena piano which plays the sequence of notes in *legato*.

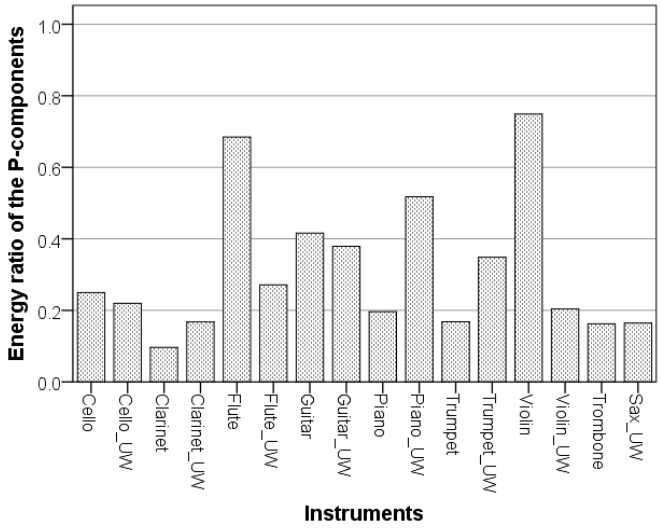

Fig. 4. Energy ratio of the P-components for different instrument samples from the instrument recognition tests MACarena and UW-CAMP. (j=15, STFT 185 ms)

B. Bass frequency extraction

The bass guitar is classified with the H-components in the "Vocals & Drums Extraction" block, which means it is attenuated or removed from the P-components. Because the bass guitar and the bass frequencies in general give more fullness to the music and the attenuation of bass frequencies may result in chopped (bass) drum sounds, the bass frequencies should be preserved in the output signal of the music pre-processing scheme. Moreover, the preference for bass sounds was also found in the music mixing preference study with CI users in [7]. To preserve the bass frequencies in the output of the music pre-processing scheme, the binary

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2428999, IEEE Transactions on Biomedical Engineering

mask for the frequency bins below a chosen cut-off frequency is uniformly set to 1, whereas the binary mask for the higher frequency bins is calculated according to their percussiveness (Figure 1). To set the cut-off frequency, eight solo bass tracks from different pop songs were passed through the music pre-processing scheme and the energy ratio of the P-components for these bass tracks was evaluated for different cut-off frequencies (Figure 5). The cut-off frequency is incorporated in the binary mask formula (9) as follows:

$$\begin{aligned} BM_{\tau,\omega} &= 1 \quad P_{\tau,\omega} > H_{\tau,\omega} \ or \quad \omega \leq \omega_{cut-off} \\ &= 0 \quad P_{\tau,\omega} \leq H_{\tau,\omega} \ and \ \omega > \omega_{cut-off} \end{aligned} \quad (17)$$

With a cut-off frequency of 400 Hz the bass guitar is mostly present in the P-components, therefore 400 Hz was chosen for the cut-off frequency in the music pre-processing scheme of Figure 1.

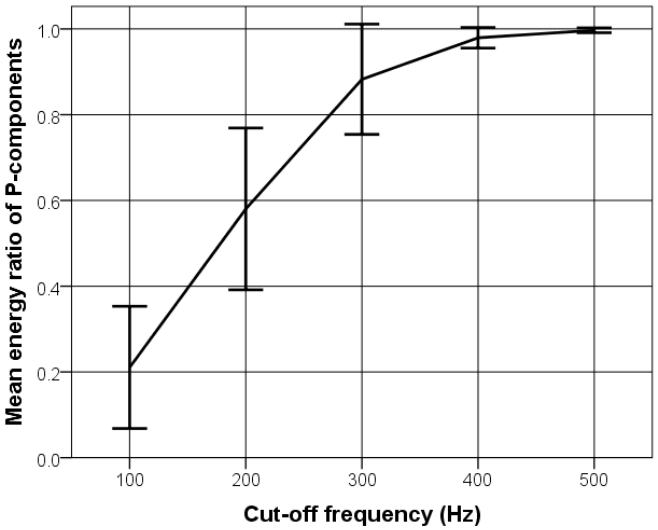

Fig. 5. Mean energy ratio of the P-components for eight solo bass guitar tracks processed with the music pre-processing scheme as a function of the cut-off frequency. Error bars represent 95% confidence interval.

C. Stereo Binary Mask

Music is widely available in stereo recordings, which means that all instruments are mixed together in two channels (left and right). Stereo recordings aim to provide the listener with the experience of a live performance in which instrument sounds are coming from the direction of the corresponding musicians on stage [29]. A typical stereo recording consists of vocals, (parts of the) drums and bass in the center and other instruments, such as guitar and piano, panned to the left or the right. This property is also used in karaoke systems to remove the vocals from the instruments. In [22], a typical voice removal algorithm for stereo recordings is described, in which the high-pass filtered left and right channel are subtracted from each other and mixed together with the low-pass filtered left and right channel. Consequently, only vocals are removed in the output, whereas the low-frequency content, such as drums and bass, is preserved. In contrast, the aim for the music pre-processing scheme is to emphasize vocals, drums and bass, which are typically mixed in the center of the stereo image. In [29] the vocals in a stereo signal are identified by comparing the spectrogram of the left channel, the right channel and the difference between left and right channel. The frequency bins for which the spectrogram of the difference is smaller than the spectrogram of the left and the right channel are classified as belonging to the vocals component. A binary mask to extract the center part of the stereo image can thus be defined as:

$$\begin{aligned} BM_{stereo} &= 1 \quad \left(\theta * {W_{\tau,\omega}}^{diff}\right) < {W_{\tau,\omega}}^{L} \ and \ \left(\theta * {W_{\tau,\omega}}^{diff}\right) < {W_{\tau,\omega}}^{R} \\ &= 0 \quad \left(\theta * {W_{\tau,\omega}}^{diff}\right) \geq {W_{\tau,\omega}}^{L} \ or \ \left(\theta * {W_{\tau,\omega}}^{diff}\right) \geq {W_{\tau,\omega}}^{R} \end{aligned} \quad (18)$$

in which θ is a tunable parameter and ${W_{\tau,\omega}}^{L}$, ${W_{\tau,\omega}}^{R}$ and ${W_{\tau,\omega}}^{diff}$ the respective spectrogram of the left channel, the right channel and the difference between the left and the right channel. Applying the binary mask on the input spectrogram results in extracting the center part of the stereo recording, which contains vocals, (part of the) drums and bass, but separation performance heavily depends on the broadness of the stereo image in the recording.

The "Vocals & Drums Extraction" block described in section II paragraph A, separates vocals and drums from the other instruments based on a single-channel input signal. For stereo recordings, the performance of the "Vocals & Drums Extraction" can be significantly improved by exploiting the inherent spatial information of the instruments. To integrate this spatial information in the optimization problem (1) an additional stereo constraint (19) has been added next to constraints (2) and (3) which limits the P-components to the center of the stereo image (cfr. formula (18)):

$$\left(\theta * {P_{\tau,\omega}}^{diff}\right) < {P_{\tau,\omega}}^{L} \ and \ \left(\theta * {P_{\tau,\omega}}^{diff}\right) < {P_{\tau,\omega}}^{R} \quad (19)$$

Consequently, the update formulae (4) and (5) are changed into:

$$P_{\tau,\omega}^{2\ (j+1)} = \frac{{\beta_{\tau,\omega}}^{(j)} \left|W_{\tau,\omega}\right|^2}{\left({\alpha_{\tau,\omega}}^{(j)} + {\beta_{\tau,\omega}}^{(j)}\right)} \quad (20)$$

$$P_{\tau,\omega}^{2\ (j+1)} = BM_{stereo} * P_{\tau,\omega}^{2\ (j+1)} \quad (21)$$

$$H_{\tau,\omega}^{2\ (j+1)} = \left|W_{\tau,\omega}\right|^2 - P_{\tau,\omega}^{2\ (j+1)} \quad (22)$$

with BM_{stereo} defined in (18) and $\alpha_{\tau,\omega}$ and $\beta_{\tau,\omega}$ defined in (6) and (7).

The SNR improvement for vocals/drums versus other instruments as a function of the stereo parameter θ from (18) is shown in Figure 6 for five different audio mixes of a typical pop song with vocals, drums, bass, guitar and piano. The audio mixes were artificially constructed with vocals, drums and bass in the center of the stereo image and different stereo panning χ ranging from 0 to 100 for piano and guitar. Panning $\chi = 0$ means no panning for piano and guitar, resulting in a mono signal, whereas panning $\chi = 100$ means complete

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2428999, IEEE Transactions on Biomedical Engineering

panning of piano and guitar respectively to the left and to the right of the stereo image. It is clear that there is no improvement in separation performance for the audio mix with panning $\chi = 0$ (mono), nor for the other audio mixes with stereo parameter θ equal to zero. Increasing the stereo parameter θ results in an improved SNR, but as shown in Figure 7 also results in distorted vocals/drums.

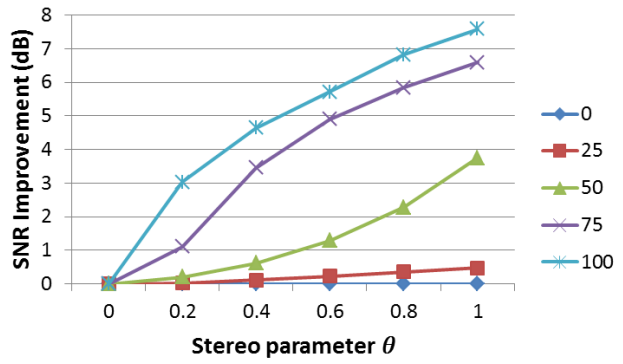

Fig. 6. SNR improvement for vocals/drums versus other instruments as a function of the stereo parameter θ from (18) for different stereo mixes with panning χ ranging from 0 to 100 for piano (panned to the left) and guitar (panned to the right).

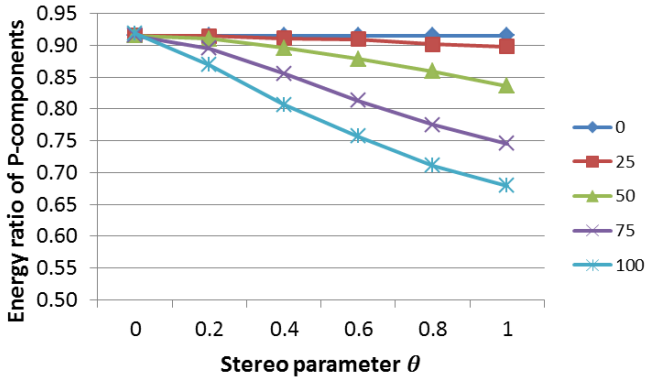

Fig. 7. Vocals/drums distortion indicated as the energy ratio of the P-components for the vocals/drums track as a function of the stereo parameter θ from (18) for different stereo mixes with panning χ ranging from 0 to 100 for piano (panned to the left) and guitar (panned to the right).

In the case that the vocals are not mixed in the center of the stereo image but panned to the left or the right, e.g. for background singers, the inclusion of the stereo parameter θ in (18) and (21) influences the vocals separation. Figure 8 shows the influence of the stereo parameter θ on the separation of the vocals represented by the energy ratio of the P-components for the vocals track. Five different audio mixes of a typical pop song were artificially constructed atypically with bass, guitar and piano in the center of the stereo image, and vocals and drums panned to left and right with panning level χ ranging from 0 to 100. Increasing the stereo parameter θ results in the removal of the vocals from the P-components depending on its panning level χ.

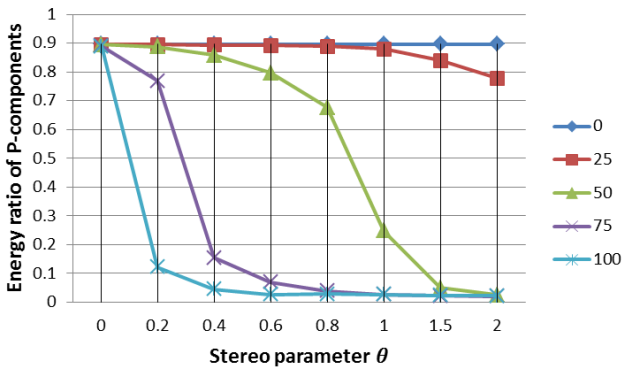

Fig. 8. Removal of the off-center vocals track (from 0 to 100) from the P-components of the music pre-processing scheme visualized as the energy ratio of the P-components for the vocals track as a function of the stereo parameter θ.

The parameter θ is determined to be 0.4, which is a good compromise between vocal distortion and instrument attenuation.

D. *Stereo music pre-processing output*

As shown in Figure 1, the obtained P-components with vocals, drums and bass are added to the H-components, which are attenuated with an adjustable parameter '*Attenuation*'. The output spectrogram after addition becomes:

$$W_{\tau,\omega}^{out} = P_{\tau,\omega} + Attenuation * H_{\tau,\omega} \quad (23)$$

The corresponding output waveform is obtained from the inverse STFT:

$$output(t) = ISTFT\left[\left(W_{\tau,\omega}^{out}\right) e^{j\angle W_{\tau,\omega}}\right] \quad (24)$$

in which the phase information from the input ($\angle W_{\tau,\omega}$) is reused. With the attenuation parameter equal to 0 dB, the output signal of the stereo music pre-processing scheme remains unaltered compared to the input signal. The final stage in the music pre-processing scheme applies a gain to the output signal as a function of the attenuation parameter to compensate for the decrease in output level due to the attenuated H-components.

III. METHODS

The stereo music pre-processing scheme with an adjustable attenuation parameter was evaluated with CI subjects. This paragraph includes a description of the sound material, the demographic and etiological information of the CI test subjects and the test procedure.

A. *Sound material*

For the perceptual evaluation of the stereo music pre-processing scheme, a selection of pop/rock songs was used

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2428999, IEEE Transactions on Biomedical Engineering

from the top fifty songs in the all-time greatest hits list of a popular radio station in Belgium (Joe FM). Representative excerpts of the songs with an average length of 27 seconds and an average dynamic range of 10.0 dB (SD = 1.5 dB) were selected. The song excerpts were rms-equalized and stored as stereo wav-files with sampling rate of 44.1 kHz. A complexity rating experiment was performed with twelve NH test subjects with no self-reported hearing deficit. The subjects were recruited with an internal advertisement, had diverse musical background and were familiar with most of the music excerpts. The fifty song excerpts were played through headphones (Beyerdynamic DT-770 pro) in a silent room and the test subjects were asked to rate the musical complexity of the song on a scale from 1 to 100 with a slider in a graphical user interface on a laptop. No further definition or information was given to the test subjects in order not to prime them in the experiment. This resulted in a ranking of the fifty songs from least complex to most complex for every subject, which was quite consistent across subjects. The average ranking over all subjects was used to compose three groups of songs, containing the 8 least complex songs, the 8 most complex songs and the 8 songs with medium complexity. These 24 excerpts were used in the experiment described in paragraph C.

B. Subjects

Seven post-lingually deafened CI subjects (all Cochlear™ Nucleus®) participated in the present study. A summary with demographic and etiological information can be found in Table I. The speech performance results - as provided by the clinic - are measured with an adaptive Speech Reception Threshold (SRT) test with LIST material and Speech Shaped Noise [30]. The subjects signed a consent form and were paid for their travel expenses. Ethical committee approval was obtained.

TABLE I
DEMOGRAPHIC AND ETIOLOGICAL INFORMATION OF SEVEN POST-LINGUALLY DEAFENED CI TEST SUBJECTS.

Subject	Age (years)	Gender	CI exp (years)	SRT (dB)	Etiology	Sound processor	Implant type
CI1	51	Male	1	-2	Colitis Ulcerosa	CP810	CI24RE(CA)
CI2	63	Male	3	+1	Progressive	CP810	CI24RE(CA)
CI3	60	Male	8	+1	Unknown	Freedom	CI24R(CS)
CI4	75	Female	5	-1	Otosclerosis	Freedom	CI24RE(CA)
CI5	55	Female	3	-3	Unknown	Freedom	CI24RE(CA)
CI6	62	Female	1	-3	Progressive	CP810	CI24RE(CA)
CI7	28	Male	10	+6	Meningitis	CP810	CI24R(CS)

C. Perceptual evaluation

For the evaluation of the stereo music pre-processing scheme, the twenty-four pop/rock songs, which were divided in three groups of eight songs with respectively low, mid and high complexity, were used. Test subjects were asked to select the preferred setting for the attenuation parameter in the range -6 dB to 30 dB, referring to the attenuation of the H-components in the output signal of the stereo music pre-processing scheme. A negative value means an amplification of the H-components. The settings were visualized on a graphical user interface with buttons numbered from 1 to 7, representing the range of the attenuation parameter from -6 dB to 30 dB (or reversed) in steps of 6 dB. The order was randomized in every trial to prevent the test subject from using the visual cue in the evaluation. The song excerpts were played continuously in a loop in free-field in a sound-treated room at a level of 60 dB(A). The experiment was performed on a laptop connected to a loudspeaker (Genelec 8020A). The songs were presented in random order and repeated three times. The actual instructions for the test subject were: (1) press the 'play' button to start listening to the first music excerpt, (2) listen to the seven modified music excerpts and identify the differences, and (3) select the music excerpt that is most enjoyable for you and press 'next' to store your preference and to move to the next music excerpt. The test subjects were allowed to take a break whenever necessary. The experiment took on average 2 to 2.5 hours per subject.

IV. RESULTS

Figure 9 shows the individual settings for the preferred attenuation of H-components for the three different groups of songs (low, mid and high complexity). In the rightmost column, the average scores from the seven subjects are indicated. All individual CI subjects preferred for every group of songs a setting with attenuated H-components which is significantly different from zero (One-Sample Wilcoxon Signed-ranks Test, $p < 0.05$). However, individual differences occur across subjects with on the one hand subject CI1 with low preferred attenuation between 0 dB and 6 dB and on the other hand subject CI4 with high preferred attenuation up to 24 dB for complex songs. Test subjects CI2, CI4, CI6 and CI7 preferred higher attenuation of H-components for songs with high complexity, whereas subjects CI1, CI3 and CI5 preferred the H-components to be attenuated irrespective of the complexity of the songs. The average preferred attenuation for each group (8 songs x 7 subjects x 3 repetitions) is analyzed with the Wilcoxon Signed-ranks Test with Bonferroni correction and is significantly higher for songs with high complexity compared to songs with low complexity ($Z = 4.71$, $p < 0.001$) or mid complexity ($Z = 2.64$, $p = 0.024$). The difference in preferred attenuation between low and mid complexity is not significant after Bonferroni correction ($Z = 2.08$, $p = 0.11$).

The preferred average setting of the attenuation parameter for the 24 song excerpts is positively correlated with the complexity of the songs as rated by the NH subjects (Pearson's $r(24) = 0.67$, $p < 0.001$) (Figure 10). For more complex songs, the preferred attenuation of the H-components is higher.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2428999, IEEE Transactions on Biomedical Engineering

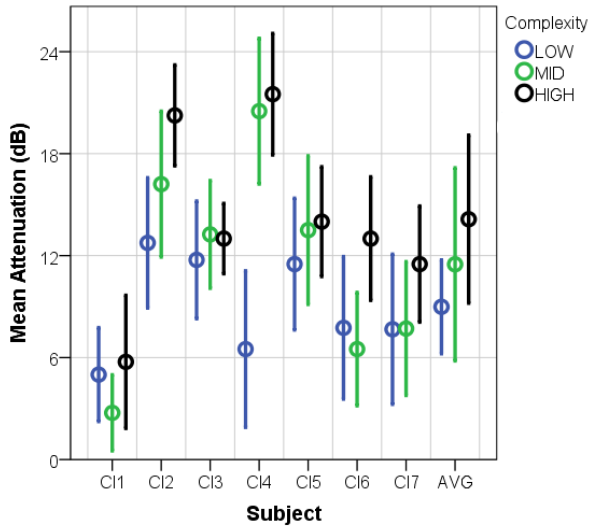

Fig. 9. Individual results for 7 CI subjects with their preferred setting for the attenuation of the H-components for 24 song excerpts with low, mid and high complexity. The average preferred setting from the seven subjects for low, mid and high complexity songs are in the rightmost column. Error bars represent 95% confidence interval.

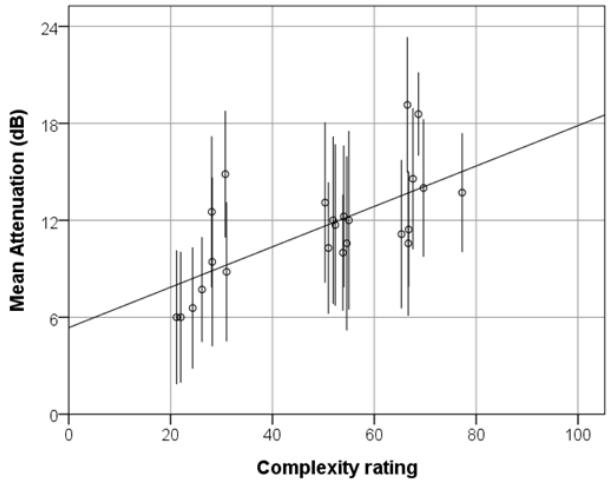

Fig. 10. Mean preferred attenuation of the H-components for the 24 song excerpts with 7 CI subjects as a function of the complexity rating given by 12 NH subjects. Error bars represent 95% confidence interval. Straight line is the linear regression ($R^2 = 0.43$).

V. Discussion

A stereo music pre-processing scheme for CI users has been described which has the potential to enhance music enjoyment by emphasizing vocals, drums and bass in complex music. The signal processing techniques used are applicable in real-time (with latency around 500 ms), which makes them interesting for music listening and/or music rehabilitation for CI users. When listening to songs from a music library, the latency is not an issue. However, for audiovisual music stimuli, the synchronization between audio and video is corrupted and should be resolved (if possible) by delaying the video signal with the (stable) latency. While attending a live performance the latency cannot be compensated, and the performance of the proposed signal processing scheme is likely to be suboptimal, unless two microphones are capturing the sound in the sweet spot of the stereo image.

The "Vocals & Drums extraction" block from Figure 1 is based on the separation of H- and P-components, which has been developed and used to improve the performance of music information retrieval tasks such as chord detection and drum detection respectively. In this paper, it is used to enhance vocals and drums in an audio mix for CI users. Different approaches for drums extraction based on harmonic/percussive sound separation are described in [20] and [31]. A comparative evaluation of various harmonic/percussive sound separation algorithms based on the anisotropic continuity in the spectrogram is described in [24]. Five different methods are analyzed and the parameter sets for each method are tuned to maximize the signal-to-interference ratio and the signal-to-distortion ratio for H or P. In [23], the same harmonic/percussive sound separation techniques are used for melody extraction by changing the STFT window length, which can be combined with the drums extraction as in [21] and Figure 3 to achieve the "Vocals & Drums Extraction" from Figure 1. The separation performance with mono signals, which is illustrated in Figure 2 for a selection of songs, can be significantly improved for stereo signals by exploiting the spatial properties of a typical stereo recording with vocals, drums and bass in the center of the stereo image and the other instruments panned to left or right. By increasing the stereo parameter θ from (18), an increase in SNR for the vocals/drums track is achieved, which is visualized in Figure 6 for different artificially constructed audio mixes. The increase in SNR is most apparent for audio mixes with instruments panned completely to left or right (panning $\chi = 100$)[1]. However, increasing the stereo parameter θ might also introduce distortions in vocals/drums, which is shown in Figure 7 with the energy ratio for the vocals/drums track. Moreover, if the vocals are not panned in the center of the audio mix, they may get distorted or completely removed in the output signal. Figure 8 shows that in an artificially constructed audio mix the vocals with e.g. panning $\chi = 75$ or $\chi = 100$ are eliminated almost entirely with stereo parameter $\theta = 0.4$. Background vocals are typically panned off-center and are thus likely to be attenuated or removed. If, on the other hand, the panning of the instruments in the stereo audio mix is not distinct or if the input signal is a mono signal, the stereo music pre-processing scheme relies only on the different representations of H- and P-components in the input spectrogram.

Clear vocals and clear rhythm/beat are described to be top factors contributing to music appreciation in CI users [3]. A study with multi-track recordings in which CI subjects (N=10) are able to adjust the different instrument levels in pop music excerpts indicates a similar preference for vocals, drums and bass in complex music [7]. The modification of the relative

[1] e.g. in old recordings from The Beatles

instrument level settings in complex music is found to be beneficial for music appreciation in CI users. The music pre-processing scheme described in this paper is able to emphasize vocal, drum and bass tracks based on mono or stereo recorded music. Songs with different complexity rating are presented to CI subjects who have been asked to determine their preferred setting for the attenuation of the H-components. By attenuating the H-components in complex music a reduction of structural complexity is achieved. The preference for a high attenuation of H-components for songs with high complexity, as found in the current study, is in agreement with the findings from [4], in which a negative correlation is found between complexity and music appreciation in CI users. The songs that are rated more complex are appreciated less and are - in the current study - preferred with higher attenuation of H-components as opposed to songs with low complexity that are already appreciated more, and thus require lower attenuation. The individual differences among subjects may relate to the experience of the CI user with listening to music. Subjects CI1 and CI7 are active music listeners, whereas the other subjects only report a more passive music listening involvement. The experienced music listener CI1, who is capable of distinguishing the different instruments in a complex song, is missing out certain instruments when attenuating the H-components, whereas a non-experienced music listener (such as subject CI4) benefits from the attenuated H-components to more easily follow the lyrics and the song in general. The sound material excerpts used in this study are pop/rock. In a future study, more different styles of music should be included together with a thorough investigation on the musical habits and music experience of the subjects and their speech and pitch performance to explain the individual differences among subjects.

VI. Conclusion

A stereo music pre-processing scheme aimed at improving music perception and appreciation in CI users has been described and evaluated. The scheme is capable of modifying the instrument level settings of music while preserving vocals, drums and bass, which has been found to be beneficial for music appreciation in CI users. The scheme has been evaluated subjectively using pop/rock song excerpts with low, medium and high complexity. On average, the preferred setting for the attenuation parameter has been found to be significantly different for the group of songs with low and high complexity. The music pre-processing scheme has the potential to improve music appreciation in CI users, in particular for complex songs. Individual differences have been observed across subjects.

Acknowledgment

We would like to thank all CI and NH subjects for participating in this music experiment, Valerie Looi and Thomas Stainsby for the interesting suggestions in defining the experiments, Hans Buyens for collecting the music tracks and Gunter Peeters for the interesting discussions on music mixing.

References

[1] P. Loizou, “Introduction to cochlear implants,” *IEEE Signal Processing Magazine*, 15, pp. 101-130, 1998.
[2] H. McDermott, “Music perception with cochlear implants: A review,” *Trends in Amplification*, 8, pp. 49-82, 2004.
[3] K. Gfeller, *et al*., “Musical backgrounds, listening habits, and aesthetic enjoyment of adult cochlear implant recipients,” *Journal of the American Academy of Audiology*, 11, pp. 390-406, 2000.
[4] K. Gfeller, *et al*., “The effects of familiarity and complexity on appraisal of complex songs by cochlear implant recipients and normal hearing adults,” *Journal of Music Therapy*, XL, pp. 78-113, 2003.
[5] Y. Kong, *et al*., “Music perception with temporal cues in acoustic and electric hearing,” *Ear and Hearing*, 25, 173-185, 2004.
[6] V. Looi, *et al*., “Music perception of cochlear implant users compared to that of hearing aid users,” *Ear and Hearing*, 29, 421-434, 2008.
[7] W. Buyens, *et al*., “Music mixing preferences of cochlear implant recipients: a pilot study,” *International Journal of Audiology*, 53(5), pp 294-301, 2014.
[8] T. Virtanen, “Sound Source Separation in Monaural Music Signals,” PhD thesis, Tampere University of Technology, 2006.
[9] J. Smith and X. Serra, “PARSHL: An analysis/synthesis program for nonharmonic sounds based on a sinusoidal representation,” *Proceedings of International Computer Music Conference*, Urbana, USA, 1987.
[10] X. Serra, “A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition,” PhD thesis, Dept. of Music, Stanford University, 1989.
[11] X. Serra, “Musical sound modeling with sinusoids plus noise,” C. Roads, S. Pope, A. Picialli, and G. D. Poli, editors, Musical Signal Processing, pp. 91–122. Swets & Zeitlinger Publishers, 1997.
[12] C. Uhle, *et al*., “Extraction of drum tracks from polyphonic music using independent subspace analysis,” *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pp. 843–847, Apr. 2003.
[13] M. Helen and T. Virtanen, “Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine,” *Proceedings EUSIPCO*, Sep. 2005.
[14] B. A. Olshausen and D. F. Field, “Sparse coding with an overcomplete basis set: A strategy employed by V1?,” *Vision Research*, 37:3311–3325, 1997.
[15] SA. Abdallah and MD. Plumbley, “Polyphonic transcription by nonnegative sparse coding of power spectra,” *Proceedings of International Conference on Music Information Retrieval*, pp. 318–325, Barcelona, Spain, October 2004.
[16] T. Virtanen. “Sound source separation using sparse coding with temporal continuity objective,” *Proceedings of International Computer Music Conference*, Singapore, 2003.
[17] A. Liutkus, *et al*., “Kernel spectrogram models for source separation,” *4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, France, 2014.
[18] A. Ozerov, *et al*., “A general flexible framework for the handling of prior information in audio source separation, ” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, Issue 4, pp 1118-1133, 2012.
[19] A. Bregman, “Auditory scene analysis,” MIT Press, Cambridge, USA, 1990.
[20] N. Ono, *et al*., “Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram,” *Proceedings EUSIPCO*, 2008.
[21] W. Buyens, *et al*., “A Harmonic/Percussive Sound Separation based Music Pre-Processing Scheme for Cochlear Implant Users,” *Proceedings EUSIPCO*, 2013.
[22] G. W. P. York, *et al*., “Teaching Real-time DSP applications (Voice Removal) with the C6711 DSK and MATLAB,” *Proceedings of the*

American Society for Engineering Education Annual Conference & Exposition, 2004.

[23] H. Tachibana, *et al.*, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," *Proceedings ICASSP*, pp. 425-428, 2010.

[24] H. Tachibana, *et al*., "Comparative evaluations of various harmonic/percussive sound separation algorithms based on anisotropic continuity of spectrogram," *Proceedings ICASSP*, pp. 465-468, 2012.

[25] O. Qazi, *et al.*, "Speech Understanding Performance of Cochlear Implant Subjects using Time-Frequency Masking based Noise Reduction," *IEEE Transactions on Biomedical Engineering*, vol. 59(5), pp. 1364-1373, 2012.

[26] S. Omran, *et al.*, "Pitch Ranking,Melody Contour and Instrument Recognition Tests Using Two Semitone Frequency Maps for Nucleus Cochlear Implants," *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[27] R. Kang, *et al*., "Development and validation of the University of Washington Clinical Assessment of Music Perception test," *Ear Hear*. 30(4):411-8, 2009.

[28] GL. Nimmons, *et al.*, "Clinical assessment of music perception in cochlear implant listeners," *Otol Neurotol*. 29(2):149-55, 2008.

[29] H. Kim, *et al*., "A Real Time Singing Voice Removal System Using DSP and Multichannel Audio Interface," *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 7, No. 2, April, 2012.

[30] A. Van Wieringen and J. Wouters, "LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands," *Int J Audiol*, 47 (6), 348-355, 2008.

[31] D. Fitzgerald, "Harmonic/Percussive Separation using Median Filtering," *13th International Conference on Digital Audio Effects (DAFX10)*, Graz, Austria, 2010.